An iterated search for influence from the future on the Large Hadron Collider

Iain Stewart

Dept. of Computing, Imperial College of Science, Technology and Medicine, London, U.K.[1]

Abstract

We analyse an iterated version of Nielsen and Ninomiya (N&N)'s proposed card game experiment to search for a specific type of backward causation on the running of the Large Hadron Collider (LHC) at CERN. We distinguish "endogenous" and "exogenous" potential causes of failure of LHC and we discover a curious "cross-talk" between their respective probabilities and occurrence timescales when N&N-style backward causation is in effect. Finally, we note a kind of "statistical cosmic censorship" preventing the influence from the future from showing up in a statistical analysis of the iterated runs.

[1] ids@doc.ic.ac.uk

1 Introduction

Nielsen and Ninomiya [1] recently proposed a particle physics model with the property that probabilities for events in the near future (say time t_1) of an “initial” state (say time t_0) depend globally on the action for complete spacetime histories, including the parts of them further in the future than t_1. The usual simplification, where we in practice consider (and sum over) only the parts of histories between spatial hypersurfaces t_0 and t_1, does not apply. This gives rise to a form of backward causation. Things like branching ratios for events here and now can depend on the ways the various alternatives can be continued to later times. Events involving Higgs production, such as the running of the Large Hadron Collider (LHC) at CERN, are the leading candidates for such effects in their model.

In [2], Nielsen and Ninomiya (henceforth N&N) propose a card game experiment where influence from the future upon the running of LHC could affect the odds of drawing cards with various instructions, such as “run LHC normally” or “shut down LHC”.

The card game version analysed by N&N in quantitative detail is a simple one-shot affair where “shut down LHC” and “run LHC normally” are present in mixing ratio p to $1 - p$. This protocol, while providing maximal contrast between the two branches, does require a huge level of “community self-discipline” which seems unlikely to be achieved in practice. Upon the drawing of “shut down LHC” the community would likely take the line that “we didn’t really mean it” and proceed with the planned runs. That very fact – a feature of the potential time evolution further into the future (running LHC) than a particular event (drawing a card) – would infect the detailed structure of the influence from the future upon the present-day enactment of the protocol.

In this paper we analyse an iterated version of the N&N protocol. The same $p : 1 - p$ pack as just described is used, but a new draw is performed at the start of each agreed time period (*e.g.* day, week, month) and the instructions “run LHC normally” or “shut down LHC” refer just to that time period. Thus at no stage is the overwhelming demand being made that LHC be permanently closed. Further, the pattern of runs and idle periods thrown up by enacting the iterated protocol is a “deliverable” in its own right, no less than the LHC run data, and available for statistical scrutiny by interested parties.

2 The relevant aspects of the N&N model

In [1], N&N discuss the possible impact on the usual classical or quantum action rules – respectively, find a history with extremal action, or sum over the exponentiated actions of all histories – of allowing an *imaginary part* in the Lagrangian specifying a history's action. They give reasons for the Higgs field to be the place one could expect an imaginary part to show up.

The mathematical effects of allowing an imaginary part in the Lagrangian are subtle and far-reaching, and according to [1] may even extend as far as determining (with probability very close to 1) a *particular* solution to the equations of motion, rather than just a recipe for computing some features of a solution from other features. Stepping back from these lofty longer-term goals, though, N&N in [2] specialize the discussion to a simplified case of their full model, where roughly identifiable "classical" trajectories (*i.e.* with probabilities rather than amplitudes attached) have been located as solutions, and the influence of the imaginary part of the Lagrangian is a simple multiplicative one, as follows.

Consider two solutions ("classical" histories) which, in the *absence* of an imaginary part in the Lagrangian, would have been assigned equal probabilities. Suppose that on one history a machine such as LHC is built and produces Higgs particles, while on the other history no such machine is built. Then, in the *presence* of the imaginary part of the Lagrangian, the relative probabilities of the two solutions are modified by an exponential in the number of Higgses:

$$\frac{P(\text{sol.}_{\text{with machine}})}{P(\text{sol.}_{\text{without machine}})} = C^{\sharp\text{Higgses}} \tag{1}$$

(Of course Higgs particles can be produced without deliberate intent, for example in the hot big bang. In [1], N&N give reasons for believing that at least in the classical approximation under consideration, the early universe may be taken as fixed across the branches, and it is legitimate to focus on the differences in *contemporary* Higgs production between one branch and another. We will accept their argument for the purposes of this paper.)

Although it is not stated explicitly in [2], we take it that the formula (1) extends in the obvious way to solutions with unequal probabilities. That is, using "uninfluenced" and "influenced" to refer to probabilities in the absence and in the presence respectively of a Lagrangian imaginary term, we have:

$$\frac{P_{\text{influenced}}(\text{sol.}_{\text{with machine}})}{P_{\text{influenced}}(\text{sol.}_{\text{without machine}})} = \frac{P_{\text{uninfluenced}}(\text{sol.}_{\text{with machine}})}{P_{\text{uninfluenced}}(\text{sol.}_{\text{without machine}})} C^{\sharp\text{Higgses}} \tag{2}$$

In (1) or (2) above, if $C < 1$ Higgs production is suppressed in processes generally, while if $C > 1$ it is enhanced. N&N consider the latter case effectively ruled out by the lack of apparent conspiracies of coincidence all around us producing more than their normal fair share of Higgs particles. Thus they (and we) focus on the case $C < 1$.

Normalization

A collection of mutually exclusive and exhaustive classical histories should be assigned probabilities which sum to 1. In standard prescriptions for assigning probabilities to classical histories, such as multiplying out the branching factors for each path on a tree of alternatives, this happens automatically provided simple local constraints are obeyed (*e.g.* that the branching factors sum to 1 at each node, in the tree of alternatives case). No final global normalization is required.

The simplified N&N model with its multiplicative influence (1) or (2) unfortunately does not assign an individually computed probability to each history. Only ratios of probabilities are prescribed. To achieve these ratios one first computes the *uninfluenced* probabilities by the usual formal or informal classical reasoning (*e.g.* by consideration of things like mixing ratios in a pack of cards, estimates of human propensity for various courses of behaviour, and so forth); and then one multiplies each history's probability by the appropriate influence factor ($C^{\sharp\text{Higgses}}$, or a generalization thereof). These *influenced but unnormalized* probabilities stand in the right ratios to one another, but they do not in general sum to 1. They must therefore be (globally) *normalized* by dividing through by their sum.

This final global normalization is perhaps the least satisfying aspect of the simplified approximate N&N model. However, it presents no technical difficulties, other than the potential complexity of computing the sum. We would be interested to know if a (presumably fully quantum) version of their model can be found which is free from the need for such normalization.

3 The iterated card game: definition

We described our iterated version of the N&N card game briefly in our introductory remarks. Here we specify it precisely, define useful notation referring to it, and clarify and motivate our assumptions (which sometimes differ from N&N's) about the various defined terms.

At the start of each time period, the intention of the human agents enacting the protocol is to draw a card from a pack, and follow the instruction printed on it, which will be either "don't run LHC this period" or "run LHC normally this period", with mixing ratio $p : 1 - p$. We assume "community self-discipline", *i.e.* the instruction drawn is followed. (Recall that the likely lack of such self-discipline in the one-shot version was our motive for moving to an iterated version of the game.)

Neither the LHC as a piece of hardware, nor the society surrounding it, offers a perfect guarantee that an intended run will actually happen. In [2], N&N use the blanket term "accident" for the thwarting of the successful running of LHC for whatever reason. The most straightforward case would be failure of the hardware, but they also give "war between the member states of CERN" as an example of the diversity of possible causes.

The potential causes of failure of LHC can broadly be divided into what might be called "endogenous" and "exogenous" kinds. An "endogenous" failure is associated with the act of trying a run, and basically refers to failure of the hardware. For simplicity we take such a failure to be severe enough to ruin LHC permanently, although one could envisage a yet finer split into categories of damage whose repair takes various lengths of time, and so forth.

An "exogenous" failure is associated with the surrounding society, and is independent of whether or not a run would be tried this time period. Indeed, it is simplest to take exogenous failure as manifesting at the very *beginning* of a period – *i.e.* the agents do not even draw a card. Analogously to our simplifying assumption about endogenous failure above, we assume an exogenous failure brings the whole protocol to an end. War between the member states of CERN would be the paradigm example.

N&N use the term "accident", assumed to have probability a, to cover all failures of LHC, but for the iterated version we will distinguish the endogenous and exogenous failures, referring to them as "accident" (to the hardware) and "breakdown" (of the whole protocol) respectively, with respective probabilities (per opportunity) a and b. (These terms are chosen for their mnemonic alphabetical adjacency and are not intended to be perfectly descriptive of the vast class of conceiveable endogenous and exogenous failures.)

Event structure in one time period

If the iterated game is still in progress as we enter a new time period (*i.e.* if no "accident" or "breakdown" has occurred earlier), exactly one of the four available events A, B, R, $\bar{R}$ will occur in this period. These stand for "accident", "breakdown", "run of LHC", and "no run of LHC" respectively. The tree of alternatives is as follows. Note that their listed probabilities are, for now, the straightforward "uninfluenced" ones got by the usual sort of probabilistic reasoning.

- First, the surrounding society impinges on the protocol. With probability b the protocol is exogenously brought to an end here ("breakdown").
  Symbol B; (uninfluenced) probability $p_B \equiv b$.

- With probability $1 - b$ the protocol continues and a card is drawn from the pack. The card has instruction "don't run LHC this period" with probability p, in which case no run is attempted. The protocol continues to the next time period.
  Symbol $\bar{R}$; (uninfluenced) probability $p_{\bar{R}} \equiv (1 - b)p$.

- Alternatively, with probability $1 - p$, the card drawn has instruction "run LHC normally this period", in which case a run is attempted. At this point we have the opportunity for an "accident" (endogenous failure) to occur, with probability a. Like "breakdown" this halts the protocol.
  Symbol A; (uninfluenced) probability $p_A \equiv (1 - b)(1 - p)a$.

- If no accident occurs the run proceeds successfully. (This is the event that will attract a multiplicative influence due to the production of Higgs particles.) The protocol continues to the next time period.
  Symbol R; (uninfluenced) probability $p_R \equiv (1 - b)(1 - p)(1 - a)$.

Classical histories of the iterated game

The iterated card game with the above event structure and stopping rules has classical histories of the form $R\bar{R}\bar{R}R\bar{R}A$ or $\bar{R}R\bar{R}\bar{R}RR\bar{R}RB$. The pattern of R and $\bar{R}$ in these examples is arbitrary; a history can be any sequence of zero or more of R or $\bar{R}$, in any order and admixture, followed by either A or B as terminating event.

It is not to be regarded as tragic that all histories end in accident or breakdown – the protocol is an expression of the sentiment "if we *can* still run LHC, *do* keep running it" (modulo the result of the card draw in each period of course). We are interested in questions such as: What happens *before* the iterated game inevitably ends? How long does the game typically last, what is the pattern of events, and which way does it end – A or B?

A history with n_R Rs, $n_{\bar{R}}$ $\bar{R}$s, n_A As, and n_B Bs (these last two being either 1 and 0, or 0 and 1 of course) has (uninfluenced) probability

$$P_{\text{uninfl.}}(\text{history}) = {p_R}^{n_R}{p_{\bar{R}}}^{n_{\bar{R}}}{p_A}^{n_A}{p_B}^{n_B} \tag{3}$$

It is automatic, and an easy exercise in algebra to verify, that the sum of the value of this expression over all the available histories is 1.

Multiplicative influence on a classical history

Let us assume that a run of LHC which lasts one time period produces some standard typical number of Higgs particles, ♯Higgses_in_one_run, and let us write σ for the expression $C^{\sharp\text{Higgses\_in\_one\_run}}$. (Note that we have $\sigma < 1$ since $C < 1$.) Then a history with n_R Rs will be subject to a multiplicative influence of σ^{n_R}.

For conceptual clarity we may as well "distribute" this influence down to the probabilities of the events which contribute to it. If for each event type $E \in \{A, B, R, \bar{R}\}$ we write ϕ_E for the *influenced* (but still unnormalized) probability of that event type (p_E being its uninfluenced probability, as listed earlier), then we have $\phi_A = p_A$, $\phi_B = p_B$, $\phi_{\bar{R}} = p_{\bar{R}}$, but $\phi_R = \sigma p_R = (1-b)(1-p)(1-a)\sigma$.

A history with n_R Rs, $n_{\bar{R}}$ $\bar{R}$s, n_A As, and n_B Bs can thus have its influenced but unnormalized probability written as

$$P_{\text{infl.,unnorm.}}(\text{history}) = {\phi_R}^{n_R}{\phi_{\bar{R}}}^{n_{\bar{R}}}{\phi_A}^{n_A}{\phi_B}^{n_B} \tag{4}$$

Of course the sum-over-histories of *this* expression will be less than 1.

Some brief notes on where our assumptions differ from N&N's

We already noted above our splitting of N&N's single failure category into separate categories of endogenous "accident" and exogenous "breakdown". For the benefit of readers steeped in N&N's notation and assumptions, we briefly note other differences between our setup and theirs.

- In their quantitative analysis N&N assume for simplicity that any branch of their tree of alternatives with Higgs particle production can be dropped altogether in their model (*i.e.* can be taken as having influenced probability zero, or negligible). In our notation this is like taking $\sigma << 1$. This would not be appropriate for the iterated protocol, since after all the time periods could be very short (minutes, seconds...)[2] and not much multiplicative influence will creep into any one period. Histories with copious Higgs production are of course still strongly suppressed, but the mathematical expression of this fact becomes, not that σ be small, but that it be raised to a high power (n_R).

- N&N take the pack mixing ratio p, and their accident probability a (which we disaggregate into a and b), to be small ($<< 1$). Their motive for choosing p small is the unacceptability of the card game to the community if p is large and the associated instruction is "shut down LHC forever". The iterated protocol does not have this problem (at least not to that stark extent!) and we can choose p more freely. As for a and b (or for N&N just a), they assume smallness to keep rough comparability with p (or just to simplify the mathematics). We prefer to drop that assumption too, since after all complex machinery *does* have a non-negligible accident rate, and societies *do* often undergo convulsions leading to the interruption of big projects (war being only one of many possible scenarios here). Thus in our analysis we do not assume p, a or b small. We are of course free to plug in small values to any expressions derived, but we derive the expressions without approximation.

- N&N perform a kind of "meta-analysis" by introducing a parameter r to represent an externally agreed probability (like a Bayesian prior) for their model

[2] At least for conceivable accelerator hardware. We leave aside whether ramping up and down the *actual* hardware of LHC on short timescales would be at all feasible!

being correct at all. (They also very modestly set $r << 1$, in effect taking the meta-analyst to be a natural skeptic of new theories.)

In this paper we are always asking: what would the world be like if the N&N model were true? Thus in effect we are always setting $r = 1$, at least for the purposes of answering that overarching question. Note, though, that with our notation one can "switch off" the model by setting $\sigma = 1$. Therefore, one *could* perform a meta-analysis by giving a prior distribution for σ (*i.e.* $1 - r$ for $\sigma = 1$, and r shared out among values of $\sigma < 1$). But this is not something that we do in this paper.

- Finally, N&N perform a cost-benefit analysis by assigning equivalent money or utility values to the outcomes of various scenarios. (Indeed, this is their main motive for having a meta-analysis variable r.) We will not do this explicitly, although from time to time we will refer to the general intuitive "goodness" or "badness" of various histories, from the viewpoint both of the scientific community (awaiting run data from LHC) and of the broader society (presumably hoping to avoid at least the more convulsive forms of the category "breakdown").

4 The iterated card game: analysis

To analyse the iterated card game we must compute the *normalized* influenced probabilities of the available histories. We first compute the normalization factor $N.F.$ by summing the influenced but unnormalized probabilities of all histories:

$$\begin{aligned} N.F. \equiv \sum_{\text{histories}} P_{\text{infl.,unnorm.}}(\text{history}) &= \sum_{0 \leq n_{R|\bar{R}} < \infty} {\phi_{R|\bar{R}}}^{n_{R|\bar{R}}} \phi_{A|B} \\ &= \frac{\phi_{A|B}}{1 - \phi_{R|\bar{R}}} \end{aligned} \tag{5}$$

(In (5) and elsewhere we abbreviate sums of probabilities or occurrence counts over alternatives by listing the alternatives in bar-separated fashion – that is, $n_{R|\bar{R}}$ means $n_R + n_{\bar{R}}$, and so forth.)

We can then divide the earlier expression (4) by $N.F.$ to obtain the normalized influenced probability for a history with n_R Rs, $n_{\bar{R}}$ $\bar{R}$s, n_A As, and n_B Bs:

$$P_{\text{infl.,norm.}}(\text{history}) \equiv \frac{P_{\text{infl.,unnorm.}}(\text{history})}{N.F.} = \frac{{\phi_R}^{n_R}{\phi_{\bar{R}}}^{n_{\bar{R}}}{\phi_A}^{n_A}{\phi_B}^{n_B}}{\frac{\phi_{A|B}}{1-\phi_{R|\bar{R}}}} \tag{6}$$

Let us specialize this rather forbidding expression to a history ending in A or B respectively – *i.e.* with $n_A = 1$, $n_B = 0$ or $n_A = 0$, $n_B = 1$ respectively.

$$P_{\text{infl.,norm.}}(\text{history ending in A}) \ = \ {\phi_R}^{n_R}{\phi_{\bar{R}}}^{n_{\bar{R}}}(\frac{\phi_A}{\phi_{A|B}}[1-\phi_{R|\bar{R}}]) \tag{7}$$

$$P_{\text{infl.,norm.}}(\text{history ending in B}) \ = \ {\phi_R}^{n_R}{\phi_{\bar{R}}}^{n_{\bar{R}}}(\frac{\phi_B}{\phi_{A|B}}[1-\phi_{R|\bar{R}}]) \tag{8}$$

This is the *same* probability distribution (assignment of probabilities to histories) as the *uninfluenced* distribution we would have obtained if, instead of p_R, $p_{\bar{R}}$, p_A, p_B in (3) above, we had used ρ_R, $\rho_{\bar{R}}$, ρ_A, ρ_B, defined as follows:

$$\rho_R \equiv \phi_R = \sigma p_R = (1-b)(1-p)(1-a)\sigma \tag{9}$$

$$\rho_{\bar{R}} \equiv \phi_{\bar{R}} = p_{\bar{R}} = (1-b)p \tag{10}$$

$$\begin{aligned}\rho_A \ &\equiv \ \frac{\phi_A}{\phi_{A|B}}[1-\rho_{R|\bar{R}}] = \frac{\phi_A}{\phi_{A|B}}[1-\phi_{R|\bar{R}}] = \frac{p_A}{p_{A|B}}[1-\sigma p_R - p_{\bar{R}}] \\ &= \ \frac{(1-b)(1-p)a}{(1-b)(1-p)a+b}[1-(1-b)(1-p)(1-a)\sigma-(1-b)p] \end{aligned} \tag{11}$$

$$\begin{aligned}\rho_B \ &\equiv \ \frac{\phi_B}{\phi_{A|B}}[1-\rho_{R|\bar{R}}] = \frac{\phi_B}{\phi_{A|B}}[1-\phi_{R|\bar{R}}] = \frac{p_B}{p_{A|B}}[1-\sigma p_R - p_{\bar{R}}] \\ &= \ \frac{b}{(1-b)(1-p)a+b}[1-(1-b)(1-p)(1-a)\sigma-(1-b)p] \end{aligned} \tag{12}$$

That is, with $\{\rho_E\}$ ($E \in \{A, B, R, \bar{R}\}$) defined as above, the normalized influenced probability of every history is given by

$$P_{\text{infl.,norm.}}(\text{history}) = {\rho_R}^{n_R}{\rho_{\bar{R}}}^{n_{\bar{R}}}{\rho_A}^{n_A}{\rho_B}^{n_B} \tag{13}$$

Thus the $\{\rho_E\}$ can be thought of as *de facto* normalized influenced probabilities for the event types E. Note, however, that for more complicated protocols – for example involving agents with memory, who do things like adjust the mixing ratio p depending on the outcomes of previous draws – the normalized influenced probability distribution is unlikely to be a mere "re-parametrization" of the uninfluenced one. That is, there will **not**, in general, exist a choice of $\{\rho_E\}$ or analogues thereof such that substituting them for the original $\{p_E\}$ or analogues thereof mimics the effects of N&N-style backward causation. We just got lucky with the simple memoryless protocol explored here!

A helpful intuitive story for obtaining ρ_R, $\rho_{\bar{R}}$, ρ_A, ρ_B from p_R, $p_{\bar{R}}$, p_A, p_B goes like this: First, the probabilities for the non-terminating events (R, $\bar{R}$) shrink under whatever multiplicative influence (if any) they are individually subject to, which happens to be an *unbalanced* (non-ratio-preserving) shrinkage, since $\rho_R < p_R$ but $\rho_{\bar{R}} = p_{\bar{R}}$. Then, the probabilities for the terminating events (A, B) grow in *balanced* (ratio-preserving) style – *i.e.* the ratio $\rho_A : \rho_B$ is the same as $p_A : p_B$ – to the extent necessary to "fill the gap" and ensure the four event probabilities sum to 1.

With the normalized influenced probability distribution established, we can proceed to study the effects of N&N-style backward causation on the statistics of the iterated card game.

The physics community perspective: how much run data can we squeeze out of LHC?

Let us first look at things from the perspective of a caricatured LHC physics community, which we define as *not* caring about goings-on in the broader society, but instead concerning itself only with how much run data can be squeezed out of LHC before the game ends (in "accident" or "breakdown").

The goal from this perspective is to maximize $\widehat{n_R}$, the expectation value of the number of periods a successful run of LHC takes place. We have a, b, σ fixed by the properties of the LHC hardware, the surrounding society, and the N&N Lagrangian respectively, but p choosable. We compute $\widehat{n_R}$:

$$\widehat{n_R} \equiv \sum_{\text{histories}} P_{\text{infl.,norm.}}(\text{history}) n_R(\text{history})$$

$$\begin{aligned} &= \frac{\rho_R}{\rho_{R|\bar{R}}} \sum_{\text{histories}} P_{\text{infl.,norm.}}(\text{history}) n_{R|\bar{R}}(\text{history}) \\ &= \frac{\rho_R}{\rho_{R|\bar{R}}} \sum_{0 \leq n_{R|\bar{R}} < \infty} [\rho_{R|\bar{R}}{}^{n_{R|\bar{R}}} \rho_{A|B}][n_{R|\bar{R}}] \\ &= \frac{\rho_R}{\rho_{A|B}} = \frac{(1-b)(1-p)(1-a)\sigma}{1-(1-b)(1-p)(1-a)\sigma-(1-b)p} \end{aligned} \tag{14}$$

By differentiating w.r.t. p we find that this expression is monotonically decreasing in p in the range $0 \leq p \leq 1$. Thus the physics community would want to set $p = 0$, that is to say, not play the card game at all but just run LHC in every period. With this choice the expression for $\widehat{n_R}$ becomes:

$$\widehat{n_R}(p=0) = \frac{(1-b)(1-a)\sigma}{1-(1-b)(1-a)\sigma} \tag{15}$$

As expected this goes to zero as $\sigma \to 0$. However, it goes to zero *more slowly* than would any $\widehat{n_R}(p > 0)$.

The broader society perspective: how long till breakdown?

We now turn to the perspective of the "broader society". Our caricature here is the opposite of that for the physics community. The broader society is defined as having no interest in the quantity of run data from LHC, nor in whether it suffers hardware failure (our category "accident"). Rather, its concern is with our category "breakdown" – this can be presumed, at least in its more convulsive forms, to be an unpleasant experience to live through.

Recall that B has an exogenously given risk rate b in our simple setup. That is, even in the absence of LHC, or more pertinently, even after an accident (A) has halted the card game, B events continue to occur with probability b per time period. (Of course, by "B" in such non-game or post-game circumstances we mean whatever sort of event in society *would* halt LHC runs had there been any. The broader society dislikes such events not for their actual or counterfactual halting of LHC, but for their potentially convulsive nature generally.) Thus we should not ask: Can we avoid B altogether? but rather: What is the expected time to the next occurrence of B? In the absence of influence from the future this is $1/b$, by elementary properties of the exponential distribution. We now compute it for the influenced probability distribution given by the $\{\rho_E\}$.

For a history ending in B, the expected time of occurrence of B (numbering the periods $1, 2, 3...$) is simply the *actual* time B occurs, *i.e.* the length of that history:

$$\widehat{t_B}(\text{history ending in B}) \;=\; t_B(\text{history}) = \text{length}(\text{history}) \tag{16}$$

For a history ending in A, we have no B within the history, but as discussed above we expect a post-game B to occur eventually. Once the accident has occurred there is no influence from the future to contend with, *i.e.* the straightforward uninfluenced risk rate b applies. Thus an occurrence of A at time t_A can be taken to herald an occurrence of B at (expected) later time $t_A + 1/b$:

$$\begin{aligned}\widehat{t_B}(\text{history ending in A}) \;&=\; t_A(\text{history}) + 1/b \\ &=\; \text{length}(\text{history}) + 1/b \end{aligned} \tag{17}$$

We can now compute $\widehat{t_B}$:

$$\begin{aligned}
\widehat{t_B} \;&\equiv\; \sum_{\text{histories}} P_{\text{infl.,norm.}}(\text{history})\widehat{t_B}(\text{history}) \\
&=\; \sum_{\text{histories ending in B}} [P_{\text{infl.,norm.}}(\text{history})][\text{length}(\text{history})] \\
&+\; \sum_{\text{histories ending in A}} [P_{\text{infl.,norm.}}(\text{history})][\text{length}(\text{history}) + 1/b] \\
&=\; \sum_{0 \le n_{R|\bar{R}} < \infty} [{\rho_{R|\bar{R}}}^{n_{R|\bar{R}}} \rho_B][n_{R|\bar{R}} + 1] + \sum_{0 \le n_{R|\bar{R}} < \infty} [{\rho_{R|\bar{R}}}^{n_{R|\bar{R}}} \rho_A][n_{R|\bar{R}} + 1 + 1/b] \\
&=\; \frac{1}{p_B} - \frac{1}{p_{A|B}} + \frac{1}{\rho_{A|B}} \\
&=\; \frac{1}{b} - \frac{1}{(1-b)(1-p)a + b} + \frac{1}{1 - (1-b)(1-p)(1-a)\sigma - (1-b)p}
\end{aligned} \tag{18}$$

(We omit the tedious algebra leading to the final “sum or difference of reciprocals” form, which seems to be the simplest possible.)

One can get an intuitive grip on (18) by setting $q \equiv 1-p$ and defining "influence constants" i_1, i_2 as follows:

$$\begin{aligned} i_1 &\equiv (1-b)a \\ i_2 &\equiv (1-b)(1-a)(1-\sigma) \end{aligned} \tag{19}$$

Then (18) becomes:

$$\widehat{t_B} = \frac{1}{b} - \frac{1}{b + i_1 q} + \frac{1}{b + (i_1 + i_2)q} \tag{20}$$

The variation of $\widehat{t_B}$ with q now becomes clear: the three reciprocals are equal when $q = 0$, but as we increase q they separate – the first staying fixed, the second and third falling, with the third (being added) always smaller than the second (being subtracted). An immediate consequence is that $\widehat{t_B}$ is at its maximum when $q = 0$ (*i.e.* when $p = 1$). However it is not in general monotonically decreasing in q in the range $0 \leq q \leq 1$: differentiating w.r.t. q shows it to reach a minimum at $q = b/\sqrt{i_1(i_1 + i_2)}$ (if that value is in the range 0..1) and start increasing again, though it never again attains its $q = 0$ value.

We see then that the broader society would prefer to set $q = 0$ ($p = 1$), that is to say, not run LHC at all. Of course $\widehat{t_B}$ is then just $1/b$. The nonmonotonicity of $\widehat{t_B}$ with q deserves further comment, however. It is best thought of as arising from a kind of "cross-talk" between, on the one hand, the effect of the influence from the future on the timescale for halting the protocol by *either* terminating event (A or B), and on the other, the impact of q on *which* terminating event it shall be. The expected time for LHC to be halted for whatever reason is

$$\begin{aligned} \widehat{t_{A|B}} &= \sum_{\text{histories}} [P_{\text{infl.,norm.}}(\text{history})][\text{length}(\text{history})] \\ &= \sum_{0 \leq n_{R|\bar{R}} < \infty} [{\rho_{R|\bar{R}}}^{n_{R|\bar{R}}} \rho_{A|B}][n_{R|\bar{R}} + 1] \\ &= \frac{1}{\rho_{A|B}} = \frac{1}{b + (i_1 + i_2)q} \end{aligned} \tag{21}$$

– and *this* is clearly monotonically decreasing in q. On the other hand, the odds ratio "ending in A" : "ending in B" is just $\rho_A : \rho_B$, which we earlier observed is the same as $p_A : p_B$, namely $(1-b)qa : b$, which becomes more favourable to A at the expense of B as q increases. This provides a "safety valve" effect: at high enough q the protocol is indeed brought to a halt very quickly, but with odds shifted in the direction of A, which (after it happens) relaxes the risk of B to its uninfluenced level (timescale $1/b$).

A numerical example

Let us put some numerical flesh on these algebraic bones. Take the time period to be a week, and set $a = b = 0.0002$ (once-per-century uninfluenced accident or breakdown timescale), $\sigma = 0.9$. If we choose $q = 0$ we have $\widehat{t_B} = \widehat{t_{A|B}} = 1/b = 5000$, the uninfluenced timescale of a century or so.

If we increase q to 0.04 (near the minimum-$\widehat{t_B}$ value, which in this example is $q \approx 0.0447$), we obtain $\widehat{t_{A|B}} \approx 238$. Thus already the likely timescale for *something* to halt the protocol has shrunk from a century to a few years. Furthermore, the odds $\rho_A : \rho_B$ are about $1 : 25$, *i.e.* the reason will almost certainly turn out to be "breakdown". These effects combine to yield $\widehat{t_B} \approx 430$.

If we make the choice $q = 1$, we get $\widehat{t_{A|B}} \approx 10$: something will halt the protocol in a matter of weeks. However, the odds $\rho_A : \rho_B$ are now $\approx 1 : 1$, so we have the "safety valve" of a 0.5 probability that the reason will turn out to be not "breakdown" but "accident". This is reflected in the value $\widehat{t_B} \approx 2510$. In effect we are tossing a fair coin and gambling on breakdown in weeks versus a century.

5 Statistical cosmic censorship

The above analysis and numerical example shows that N&N-style backward causation can be pretty powerful stuff. It can greatly enhance the likelihood of a quick occurrence of what we would normally regard as remote contingencies. It is natural to ask: what are the prospects for *empirical* study of the influence from the future? Would we know it when we see it?

If we enact the iterated card game protocol and quickly experience an event of the sort we have called "accident" or "breakdown", we are left in the awkward situation of not really knowing (as opposed to estimating) the *uninfluenced* probabilities of these contingencies, which depend on the nature of complex hardware and a complex society. People of goodwill will disagree in their estimates. Those whose estimates are on the high side will simply give a resigned shrug. “That’s life”, they will say.

There *is*, however, one feature of the protocol which has a clear-cut uninfluenced probabilistic structure: the drawing of a card in each time period. Everyone can agree that the uninfluenced probability of drawing the instruction “don’t run LHC this period” is p. And as we mentioned in our introductory remarks, the pattern of runs and idle periods is the protocol’s “deliverable”, available for statistical scrutiny by interested parties. What can they conclude if they perform a statistical analysis? Can they detect the hand of N&N-style backward causation?

A card draw occurs part-way through its time period and there is not a one-one mapping between results of draws and event labels. A drawing of “run LHC normally this period” (which we will abbreviate “yes”, Y) happens in events A and R. A drawing of “don’t run LHC this period” (which we will abbreviate “no”, N) happens only in event $\bar{R}$. No draw at all occurs in event B since exogenous halting is assumed to occur at the beginning of a time period. Hence a draw pattern (of Y and N) is got from a regular event history by dropping B, changing A and R into Y, and changing $\bar{R}$ into N.

We can conveniently quantify the strength of the influence from the future on the pattern of card draws by defining the *discrepancy* D between the actual number of drawings of “no” and the number expected purely on the basis of multiplying the total number of draws by the “no” mixing ratio p. That is:

$$D \equiv n_N - (n_{Y|N})p \tag{22}$$

Those using uninfluenced probabilistic reasoning will assign this an expectation value $\widehat{D} = 0$ – although of course they anticipate that the actual (outcome) discrepancy will suffer binomial-style fluctuation of order $\sqrt{(n_{Y|N})pq}$ about its zero expected value.

We compute the expected discrepancy under the *influenced* probability distribution given by the $\{\rho_E\}$:

$$\begin{aligned}\widehat{D} &\equiv \sum_{\text{histories}} P_{\text{infl.,norm.}}(\text{history})D(\text{history}) \\ &= \sum_{\text{histories}} P_{\text{infl.,norm.}}(\text{history})[n_N - (n_{Y|N})p](\text{history})\end{aligned}$$

$$\begin{aligned}
&= \sum_{\text{histories}} P_{\text{infl.,norm.}}(\text{history})[n_{\bar{R}} - (n_{A|R|\bar{R}})p](\text{history}) \\
&= \sum_{0\leq n_{R|\bar{R}}<\infty} [{\rho_{R|\bar{R}}}^{n_{R|\bar{R}}}\rho_A][\frac{\rho_{\bar{R}}}{\rho_{R|\bar{R}}} n_{R|\bar{R}} - (n_{R|\bar{R}}+1)p] \\
&+ \sum_{0\leq n_{R|\bar{R}}<\infty} [{\rho_{R|\bar{R}}}^{n_{R|\bar{R}}}\rho_B][\frac{\rho_{\bar{R}}}{\rho_{R|\bar{R}}} n_{R|\bar{R}} - (n_{R|\bar{R}})p] \\
&= [\ p\ ]\ [\ \frac{p_B}{p_{A|B}}\ ]\ [\ 1 - \frac{p_{A|B}}{\rho_{A|B}}\ ]
\end{aligned} \tag{23}$$

(We again omit the tedious algebra leading to the final form.) $\widehat{D}$ is thus the product of three factors each manifestly in the range 0..1, and is therefore in the range 0..1 also.[3] In other words, the influence from the future shifts $\widehat{D}$ by **less than a single card** from its uninfluenced value – a shift lost in the order-$\sqrt{(n_{Y|N})pq}$ fluctuation in the actual (outcome) discrepancy. Statistical analysis of the card draws will *not* reveal the hand of N&N-style backward causation.

Hence we find ourselves in a startling epistemic situation if we try to do statistical analysis in an N&N world. The "epistemically opaque" aspects of the enactment of the protocol, such as the timescale for the one-off event of being forcibly halted, are potentially greatly influenced by N&N-style backward causation; while the one clearly epistemically accessible aspect – the drawing of a card, which has a straightforward probability structure controlled by the mix of cards in the pack, and which is a repeated rather than one-off event, allowing the accumulation of statistics – is not *detectably* influenced at all! It is thus peculiarly difficult (at least with this protocol) for the inhabitants of a world subject to N&N-style backward causation to gather evidence revealing their predicament. This situation might be called "statistical cosmic censorship".

We close by conjecturing that statistical cosmic censorship may be a generic, or at least common, feature of models with N&N-style backward causation – and,

[3] It may be helpful to make judicious use of $q \equiv 1-p$ and the "influence constants" i_1, i_2 defined earlier:

$$\widehat{D} = [\ p\ ]\ [\ \frac{b}{b+i_1 q}\ ]\ [\ \frac{i_2 q}{b+(i_1+i_2)q}\ ] \tag{24}$$

It is then immediate that each factor is in the range 0..1.

perhaps, of theories with other types of non-standard casual structure too. We would welcome efforts to define the syndrome more precisely, and to explore what reliable knowledge can and cannot be acquired by agents enacting experimental protocols of their choice in a world where the future, as well as the past, informs their actions.

Acknowledgements

I would like to thank Paulo Pires-Pacheco for drawing my attention to [2], and for robust discussion of it and other things; and Adilah Hussein and Shashank Virmani, for discussions on causality in physics and much else besides.

References

[1] H. B. Nielsen and M. Ninomiya, "Future Dependent Initial Conditions from Imaginary Part in Lagrangian", Proceedings of the 9th Workshop "What Comes Beyond the Standard Models", Bled, 16 - 26 September 2006, DMFA Zaloznistvo, Ljubljana, arXiv:hep-ph/0612032.

[2] H. B. Nielsen and M. Ninomiya, "Search for Effect of Influence from Future in Large Hadron Collider", arXiv:0707.1919.